Department of Mathematics and Systems Analysis

Finite Element Methods for Time-Harmonic Wave Equations

Antti Hannukainen

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 88/2011

Finite Element Methods for Time-Harmonic Wave Equations

Antti Hannukainen

Doctoral dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the School of Science for public examination and debate in Auditorium G at the Aalto University School of Science (Espoo, Finland) on the 21th of October 2011 at 12:00 o'clock.

Aalto University
School of Science
Department of Mathematics and Systems Analysis

Supervisor
Prof. Rolf Stenberg

Instructor
Prof. Rolf Stenberg

Preliminary examiners
Prof. Joseph Pasciak, Texas A&M University, USA
Prof. Jay Gopalakrishnan, Portland State University, USA

Opponent
Prof. Ragnar Winther, University of Oslo, Norway

Aalto University publication series
DOCTORAL DISSERTATIONS 88/2011

ISBN 978-952-60-4297-8 (pdf)
ISBN 978-952-60-4296-1 (printed)
ISSN-L 1799-4934
ISSN 1799-4942 (pdf)
ISSN 1799-4934 (printed)

Aalto Print
Helsinki 2011

Finland

The dissertation can be read at http://lib.tkk.fi/Diss/

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Antti Hannukainen

Name of the doctoral dissertation
Finite Element Methods for Time-Harmonic Wave Equations

Publisher School of Science

Unit Department of Mathematics and Systems Analysis

Series Aalto University publication series DOCTORAL DISSERTATIONS 88/2011

Field of research Mathematics

Manuscript submitted 14 June 2011 **Manuscript revised** 16 August 2011

Date of the defence 21 October 2011 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

This thesis concerns the numerical simulation of time-harmonic wave equations using the finite element method. The main difficulties in solving wave equations are the large number of unknowns and the solution of the resulting linear system. The focus of the research is in preconditioned iterative methods for solving the linear system and in the validation of the result with a posteriori error estimation. Two different solution strategies for solving the Helmholtz equation, a domain decomposition method and a preconditioned GMRES method are studied. In addition, an a posterior error estimate for the Maxwell's equations is presented.

The presented domain decomposition method is based on the hybridized mixed Helmholtz equation and using a high-order, tensorial eigenbasis. The efficiency of this method is demonstrated by numerical examples. As the first step towards the mathematical analysis of the domain decomposition method, preconditioners for mixed systems are studied. This leads to a new preconditioner for the mixed Poisson problem, which allows any preconditioned for the first order finite element discretization of the Poisson problem to be used with iterative methods for the Schur complement problem.

Solving the linear systems arising from the first order finite element discretization of the Helmholtz equation using the GMRES method with a Laplace, an inexact Laplace, or a two-level preconditioner is discussed. The convergence properties of the preconditioned GMRES method are analyzed by using a convergence criterion based on the field of values. A functional type a posterior error estimate is derived for simplifications of the Maxwell's equations. This estimate gives computable, guaranteed upper bounds for the discretization error.

Keywords Finite element method, time-harmonic wave equations, Helmholtz equation, fast solution methods

ISBN (printed) 978-952-60-4296-1 **ISBN (pdf)** 978-952-60-4297-8

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Espoo **Location of printing** Helsinki **Year** 2011

Pages 139 **The dissertation can be read at** http://lib.tkk.fi/Diss/

A? Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Antti Hannukainen

Väitöskirjan nimi
Finite Element Methods for Time-Harmonic Wave Equations

Julkaisija Perustieteiden Korkeakoulu

Yksikkö Matematiikan ja Systeemianalyysin laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 88/2011

Tutkimusala Matematiikka

Käsikirjoituksen pvm 14.06.2011 **Korjatun käsikirjoituksen pvm** 16.08.2011

Väitöspäivä 21.10.2011 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Väitöskirjassa tutkitaan aika-harmonisten aaltoyhtälöiden ratkaisemista elementtimenetelmän avulla. Suurimmat haasteet aaltoyhtälöiden numeerisessa ratkaisemisessa ovat vaadittavien vapausasteiden suuri määrä sekä syntyvän yhtälöryhmän ratkaiseminen. Työssä keskitytään Helmholtzin yhälön iteratiiviseen ratkaisemiseen sekä Maxwellin yhtälöiden elementtiratkaisun tarkkuuden arviointiin a posteriori virhearvion avulla. Helmholtzin yhtälön ratkaisemisessa käytetään alueenhajoitusmenetelmä sekä pohjustettua GMRES iteraatiota.

Alueenhajoitusmenetelmä perustuu Helmholtzin yhtälön hybridisoituun sekaelementtiformulaatioon ja korkea-asteisen, tensorirakenteisen ominaisfunktiokannan käyttöön. Menetelmän tehokkuus osoitetaan numeerisin kokein. Alueenhajoitusmenetelmää analysoitaessa syntyi uusi pohjustin Poissonin yhtälön sekaelementtiformulaatiolle, joka mahdollistaa Poissonin yhtälön elementtiaproksimaatioille aiemmin kehitettyjen pohjustimien soveltamisen tässä yhteydessä.

Tämän lisäksi väitöskirjassa tutkitaan Helmholtzin yhtälön elementtiaproksimaatiosta syntyvän yhtälöryhmän iteratiiviseen ratkaisemiseen soveltuvia pohjustimia. Tutkimuksen kohteena ovat kahden tason pohjustin, Laplacen pohjustin, sekä epätarkkaa Laplacen pohjustin. Työssä tutkitaan näillä mentelmillä pohjustetun GMRES iteraation suppenemista ratkaistavaan matriisiin liittyvän ``field of values''-joukon avulla.

Avainsanat Elementtimenetelmä, aika-harmoniset aaltoyhtälöt, Helmholtzin yhtälö, nopeat ratkaisumenetelmät

ISBN (painettu) 978-952-60-4296-1 **ISBN (pdf)** 978-952-60-4297-8

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Espoo **Painopaikka** Helsinki **Vuosi** 2011

Sivumäärä 139 **Luettavissa verkossa osoitteessa** http://lib.tkk.fi/Diss/

Preface

This thesis has been written at the Institute of Mathematics and Systems Analysis of the Aalto university during 2007-2011. I am grateful for the financial support that I have received from Finnish Academy of Science and Letters. Part of the work has been supported by the Academy of Finland grant 133174.

Many people have helped me during these four years. First and foremost, my supervisor Prof. Rolf Stenberg has always had time for my questions on the finite element method. My master thesis supervisor Prof. Sergey Korotov gave me an excellent start to my PhD studies. My work with him is the main motivation behind Publication IV. I would also like to thank the preliminary examiners of my thesis, Prof. Joseph Pasciak and Prof. Jay Gopalakrishnan, for their time and their valuable comments.

I was fortunate to have several possibilities to discuss and to exchange ideas with other researchers. My visits to the research group of Prof. Joachim Schöberl at RWTH Aachen had a significant impact on my research. Our joint work on Publication III led me to study the Helmholtz equation. The desire to analyze the method we presented, was the main motivation behind Publications I and II. I have also been a frequent visitor of Prof. Michal Křížek at the Institute of Mathematics of the Academy of Sciences of the Czech Republic in Prague. Discussions with him have deepened my understanding of the geometric properties of finite element methods.

The Institute of Mathematics and Systems Analysis provided an excellent place for my research. My colleagues at the research group on finite element methods were always enthusiastic to exchange new mathematical ideas. Besides mathematics, we have spent many delightful evenings together, enjoying good food and life. I hope that these friendships will last for a lifetime.

My family, Elina's family, relatives and friends have shown a great interest in my research. The have challenged me to explain my work from a different perspective. Their interest has meant a lot to me. By far the most important support to my work came from my wife Elina, in the form of her love and caring. In the moments of doubt, she always had faith in my ability to finnish this project - thank you for everything.

Otaniemi, September 27, 2011,

Antti Hannukainen

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I A. Hannukainen. Field of values analysis of preconditioners for the Helmholtz equation in lossy media. *36 pages, arXiv:1106.0424*, June 2011.

II A. Hannukainen. Continuous preconditioners for the mixed Poisson problem. *19 pages, BIT Numer. Math, published electronically, http://dx.doi.org/10.1007/s10543-011-0346-0* , July 2011.

III A. Hannukainen, M. Huber, J. Schöberl. A mixed hybrid finite element method for the Helmholtz equation. *Journal of Modern Optics*, 58, Nos. 5-6, 10-20, p.424-437, March 2010.

IV A. Hannukainen. Functional Type A Posteriori Error Estimates for Maxwell's Equations. In *Numerical Mathematics and Advanced Applications, Proceedings of ENUMATH 2007*, p.41-48, 2008.

Author's Contribution

Publication I: "Field of values analysis of preconditioners for the Helmholtz equation in lossy media"

This article represents independent research of the author.

Publication II: "Continuous preconditioners for the mixed Poisson problem"

This article represents independent research of the author.

Publication III: "A mixed hybrid finite element method for the Helmholtz equation"

The results in this article were obtained in collaboration of the three authors. The article was written in tight collaboration by the first two authors.

Publication IV: "Functional Type A Posteriori Error Estimates for Maxwell's Equations"

This article represents independent research of the author.

1. Introduction

Waves in their various forms are all around us. Our speech is an ensemble of different kinds of sound waves. When we talk over the mobile phone, the conversation is sent over the air as an electromagnetic wave. Naturally, problems related to waves are frequently faced in the design of new devices. For example, how should the antenna of the mobile phone be designed for the best signal strength? Or what kind of materials should be chosen to the concert hall for the best possible musical experience?

To make good design decisions related to practically any engineering problem, numerical simulations are required. For example, before a physical prototype of a new mobile phone antenna is built, it has long before existed as an virtual model in memory of the computer. The virtual model allows the properties of the new design to be studied before costly prototypes are built. For example, the transmission properties of the mobile phone antenna can be simulated beforehand and the best design can be chosen for further development.

Studying the properties of any device using a computer requires modeling and numerical simulation steps. In the modeling step, the designer identifies the physical phenomena relevant for the device and defines a suitable set of equations capturing this phenomena. For example, the antenna might be modeled using time-harmonic Maxwell's equations. After the model is fixed, numerical simulations are used to obtain information on the properties of the device.

In most engineering problems, the numerical simulation step consist of a discretization of a partial differential equation (PDE) and a solution of the resulting system of equations. The system of equations can be either linear or non-linear. The non-linear equations are solved using iterative methods, which leads to solving series of linear problems. After a solution is obtained, its quality has to be assessed to guarantee good design

decisions.

The numerical simulation step is especially difficult for wave-type equations, see [32]. The two main issues are difficulties with discretization and with the solution of the resulting linear systems. The first of these difficulties is related to the large number of unknowns required for the solution of the PDE. The engineering explanation for this phenomenon is the Nyquist sampling theorem, which states that a certain number of sampling points per wavelength is required to resolve a wave. For example, when solving the Helmholtz equation, the engineering rule of thumb is that the minimum number of elements required for each wavelength is 10-12. If this rule is obeyed, the discretization step typically produces very large systems. For example, a simulation of the acoustic properties of a concert hall with the dimensions $20 \times 20 \times 50$ meters is computationally a very large scale problem. As the wave length of human speech is roughly in the range from four meters to thirty centimeters, following the engineering rule of thumb leads to approximately 70 million unknowns. Unfortunately, the situation is even worse. A detailed mathematical analysis reveals that the required number of unknowns grows in powers of the wave length (see [17, 18]).

Large number of unknowns is not necessarily untolerable. This is the case for mechanical engineering problems, where the largest models can have millions of unknowns, usually arising from very complex models coupling several different physical phenomena. The main difference to wave equations is the availability of efficient solvers for linear systems. Such solvers are available for many mechanical engineering problems, but unfortunately not for the linear systems arising from time-harmonic wave problems. As we will see, the work required to solve a time-harmonic wave problem grows as a function of the frequency of the modeled field

In this thesis, the main focus is in the numerical simulation step for wave problems. All of the research has been made in the context of finite element methods. The motivation of the work comes from the numerical simulation of time-harmonic Maxwell's equations. The Maxwell's equations contain two difficult phenomenon, the large kernel of the curl-operator and the wave nature of the solution. Our focus will be solely on the difficulties related to the wave nature of the solution. Hence, we have studied the Helmholtz equation, which is a simple time-harmonic wave equation. As same basic principles, and even the same governing equations, apply for numerical simulation of different kinds of time-harmonic

wave equations, the obtained results can be applied in several different fields.

The main contributions of the thesis are in the iterative solution of linear systems arising from the discretization of time-harmonic wave equations. Two different solution strategies have been studied, a domain decomposition based method and a preconditioned GMRES method. In analyzing the domain decomposition based method, preconditioners for mixed systems were studied, which led to a new preconditioner. In addition to the solution of the linear system, a minor contribution was made in the assessment of the quality of the solution to different simplifications of the Maxwell equations.

The organization of this thesis is as follows. First, we will discuss the Maxwell's equations and show what kind of wave equation is obtained from this system. Then we will introduce the model problem. With the help of the model problem, we will demonstrate the difficulties faced in numerical simulation of wave equations. Finally, a discussion of the articles, which form the main part of the thesis, is given.

2. Maxwell's equations

The Maxwell's equations

$$-\partial_t \boldsymbol{D} + \operatorname{curl}(\boldsymbol{H}) = \boldsymbol{J} \tag{2.1}$$

$$\partial_t \boldsymbol{B} + \operatorname{curl}(\boldsymbol{E}) = 0 \tag{2.2}$$

$$\operatorname{div}(\boldsymbol{B}) = 0 \tag{2.3}$$

$$\operatorname{div}(\boldsymbol{D}) = \varrho. \tag{2.4}$$

are the governing equations of electromagnetic phenomenon. They describe the interactions between currents $\boldsymbol{J}$, electric charges ϱ, electric field $\boldsymbol{E}$, electric field density $\boldsymbol{D}$, magnetic field $\boldsymbol{H}$, and magnetic field density $\boldsymbol{B}$. The Maxwell's equations are present in several different engineering applications, ranging from the design of antennas to the design of electric motors.

The Maxwell's equations are coupled with the constitutive relations

$$\boldsymbol{D} = \boldsymbol{D}\left(\boldsymbol{E}\right) \qquad \text{and} \qquad \boldsymbol{B} = \boldsymbol{B}\left(\boldsymbol{H}\right), \tag{2.5}$$

which are material dependent. The constitutive relations can be very complicated, for example the relation between magnetic field and magnetic field density in steel is highly non-linear and depends on the past values of the fields.

In many engineering problems, the electromagnetic properties of the media are modeled with sufficient accuracy by assuming the relationships (2.5) linear. In linear material, the simplest form of constitutive equations is

$$\boldsymbol{D} = \epsilon \boldsymbol{E} \qquad \boldsymbol{B} = \mu \boldsymbol{H}. \tag{2.6}$$

where $\epsilon, \mu \in \mathbb{R}$and $\epsilon > 0, \mu > 0$.

As we know from basic physics, the electric field causes charged particles to move. For Maxwell's equations this behavior is modeled by generalized Ohm's law relating the current density to the electric field. In linear material, the generalized Ohm's law has the form

$$\boldsymbol{J} = \sigma \boldsymbol{E} + \boldsymbol{J}_S, \tag{2.7}$$

in which $\sigma \in \mathbb{R}$ is the called conductivity and $\boldsymbol{J}_S$ the imposed current density. The imposed current density is a useful modeling tool. For example, a current loop can be modeled as an impose current density.

The Maxwell's equations are a detailed model of the physical phenomena related to electromagnetic fields. In many cases, such a detailed model is not required and a simpler set of equations describing the relevant phenomena with sufficient accuracy is derived from the full Maxwell system. For example, when devices operating at the low-frequency range, such as transformers, generators, or electric motors are studied, a much simpler eddy-current model is usually applied, see [1].

The simplest wave equation derived from the Maxwell's equations is the time-harmonic vector wave equation. This equation is derived under the assumptions that all materials are linear and all excitations are sinusoidal, which is often the case in practical applications. Under these assumptions, all fields are also sinusoidal and the time dependency can be described as

$$\boldsymbol{E}(\boldsymbol{x}, t) = \Re\left(\boldsymbol{E}(\boldsymbol{x})\, e^{i\omega t}\right). \tag{2.8}$$

Such time dependency allows elimination of all time derivatives and leads to the time-harmonic Maxwell's equations in the frequency domain. The equations (2.1) and (2.2) take the form

$$-i\omega \boldsymbol{D} + \operatorname{curl}(\boldsymbol{H}) = \boldsymbol{J} \tag{2.9}$$

$$i\omega \boldsymbol{B} + \operatorname{curl}(\boldsymbol{E}) = 0. \tag{2.10}$$

In materials satisfying constitutive relations (2.6), electric and magnetic field densities can be eliminated. The term $\sigma \boldsymbol{E}$ in Ohm' s law (2.7) is taken into account by introducing complex permittivity

$$\tilde{\epsilon} = \epsilon \left(1 + \frac{i\sigma}{\omega\epsilon}\right). \tag{2.11}$$

Using this notation, leads to equations

$$-i\omega\tilde{\epsilon}\boldsymbol{E} + \text{curl}(\boldsymbol{H}) = \boldsymbol{J}_s \tag{2.12}$$

$$i\omega\mu\boldsymbol{H} + \text{curl}(\boldsymbol{E}) = 0. \tag{2.13}$$

These equations can be further simplified by eliminating either $\boldsymbol{E}$ or $\boldsymbol{H}$ field. Eliminating the $\boldsymbol{H}$ field leads to the vector-wave equation

$$\text{curl}\big(\mu^{-1}\text{curl}(\boldsymbol{E})\big) - \kappa^2\boldsymbol{E} = \boldsymbol{J}_s \tag{2.14}$$

where the wave number $\kappa^2 = \omega^2\tilde{\epsilon}$. This equation represents the simplest wave-type equation derived from the system (2.1)-(2.4).

In mathematical analysis, problem (2.14) is interpreted in its weak form. For this purpose, we will assume that the problem is posed in a simply connected domain Ω with perfectly electrically conducting (PEC) surface such that

$$\boldsymbol{n} \times \boldsymbol{E} = 0 \text{ on } \partial\Omega.$$

Under these assumptions, the weak form of the problem (2.14) is: Find $\boldsymbol{E} \in H_0(\text{curl}; \Omega)$ such that

$$(\mu^{-1}\text{curl}(\boldsymbol{E}), \text{curl}(\boldsymbol{u})) - \kappa^2(\boldsymbol{E}, \boldsymbol{u}) = (\boldsymbol{J}_s, \boldsymbol{u}) \quad \forall \boldsymbol{u} \in H_0(\text{curl}; \Omega). \tag{2.15}$$

Here, $(\cdot, \cdot)$ is the standard $L^2(\Omega)$-inner product and the space $H_0(\text{curl}; \Omega)$ is defined as

$$H_0(\text{curl}; \Omega) = \left\{ \boldsymbol{u} \in L^2(\Omega) \mid \|\boldsymbol{u}\|_{curl} < \infty \text{ and } \boldsymbol{n} \times \boldsymbol{u} = 0 \text{ on } \partial\Omega \right\}.$$

where the norm $\|\boldsymbol{u}\|_{curl}$ is defined as

$$\|\boldsymbol{u}\|_{curl}^2 := \|\boldsymbol{u}\|_0^2 + \|\text{curl}(\boldsymbol{u})\|_0^2.$$

This function space is very natural for the Maxwell's equations. The functions from $H_0(\text{curl}; \Omega)$ have continuous tangential components over mate-

rial interfaces, which is also the case for the electric field $\boldsymbol{E}$.

In the first finite element solvers for Maxwell's equations, the weak problem (2.15) was solved using $[H^1(\Omega)]^3$-conforming finite element methods. This approach led to difficulties in the numerical simulation step, for example spurious modes appeared in eigenvalue computations of non-convex domains, [24, 1]. These difficulties were due to the $[H^1(\Omega)]^3$ -conforming finite element methods being incorrect for Maxwell's equations. Nowdays, these difficulties have been overcome, e.g., by using the $H(\mathrm{curl};\Omega)$-conforming methods in finite element simulations.

From mathematical point of view, the Maxwell equations contain two interesting phenomenon, the large kernel of the curl-operator and the wave type behavior of the solution. As in [24], these two phenomenon can be isolated in the mathematical analysis by using the Helmholtz decomposition of vector fields.

The Helmholtz decomposition splits a vector field into two parts, $\boldsymbol{u} = \boldsymbol{u}_0 + \nabla p$. Several alternative decompositions with different kind of requirements for the fields $\boldsymbol{u}_0$ and p exist. Here, we will use the decomposition such that

$$(\boldsymbol{u}_0, \nabla\xi) = 0 \quad \forall \xi \in H_0^1(\Omega).$$

The Helmholtz decomposition divides the space $H_0(\mathrm{curl};\Omega)$ into two parts

$$H_0(\mathrm{curl};\Omega) = X_0 \oplus \nabla H_0^1(\Omega),$$

where the space X_0 is defined as

$$X_0 := \{\ \boldsymbol{u} \in H_0(\mathrm{curl};\Omega) \mid (\boldsymbol{u}, \nabla p) = 0 \quad \forall p \in H_0^1(\Omega)\ \}. \tag{2.16}$$

Using the Helmholtz decomposition, the vector wave equation is split into two parts: find $\boldsymbol{E}_0 \in X_0$ and $p \in H_0^1(\Omega)$ such that

$$-\kappa^2(\nabla p, \nabla\xi) = (\boldsymbol{J}_s, \nabla\xi) \quad \forall \xi \in H_0^1(\Omega) \tag{2.17}$$

$$(\mu^{-1}\mathrm{curl}(\boldsymbol{E}_0), \mathrm{curl}(\boldsymbol{u}_0)) - \kappa^2(\boldsymbol{E}_0, \boldsymbol{u}_0) = (\boldsymbol{J}_s, \boldsymbol{u}_0) \quad \forall \boldsymbol{u}_0 \in X_0 \tag{2.18}$$

The first of these equations is the Poisson problem related to the kernel of the curl-operator. It does not exhibit any wave-type behavior. The second equation has the structure typical for time-harmonic wave equations: a differential operator and a lower order shift term. As the main focus of

the thesis is in time-harmonic wave equations, this structure motivates us to consider a simpler model problem with the same properties, namely the Helmholtz equation.

2.1 The model problem

The Helmholtz equation is a prototypical time-harmonic wave equation. It arises in several physical situations, for example in the simulation of sound waves. From mathematical point of view, the Helmholtz equation has a similar structure with the wave-part of the time-harmonic vector wave equation (2.18), namely a differential operator with a lower order sifth term. Naturally, the curl-operator present in vector wave equation is much more complicated than the Laplace operator, but similar properties are shared by the two. For example, the Poincare inequality is valid in both spaces $H_0^1(\Omega)$ and X_0.

The Helmholtz equation is: Find u such that

$$-\Delta u - (\kappa^2 - \mathrm{i}\sigma)u = f \quad \text{in } \Omega. \tag{2.19}$$

for simplicity, we will consider either the homogenous Dirichlet boundary condition

$$u = 0 \text{ on } \partial\Omega \tag{2.20}$$

or the absorbing boundary condition

$$\frac{\partial u}{\partial \boldsymbol{n}} - \mathrm{i}\kappa u = g \text{ on } \partial\Omega. \tag{2.21}$$

The domain $\Omega \subset \mathbb{R}^d, d = 2, 3$. The functions f and g are from the spaces $L^2(\Omega)$ and $L^2(\partial\Omega)$, respectively. The parameters κ and σ are both real valued $\kappa, \sigma \in \mathbb{R}$. The parameter κ is always positive $\kappa > 0$. The parameter σ is positive, $\sigma > 0$, for the Dirichlet boundary condition case (2.20) and zero, $\sigma = 0$, for the absorbing boundary case (2.21).

The shape of the domain Ω has an effect on the properties of the weak solution u. For simplicity, we will assume that domain Ω is convex, leading to $H^2(\Omega)$-regularity of the weak solution when homogenous Dirichlet boundary conditions are imposed, see Publication I and [6, 15]. The regularity is an important property of the solution, affecting the convergence of discretization methods for the PDE's.

The weak form of the problem (2.19) with homogenous Dirichlet boundary conditions (2.20) is: Find $u \in H_0^1(\Omega)$ such that

$$(\nabla u, \nabla v) - \kappa^2 (u, v) + \mathrm{i}\sigma(u, v) = (f, v) \quad \forall v \in H_0^1(\Omega) \tag{2.22}$$

and with the absorbing boundary conditions : Find $u \in H^1(\Omega)$ such that

$$(\nabla u, \nabla v) - \kappa^2 (u, v) - \mathrm{i}\kappa \langle u, v \rangle_{\partial\Omega} = (f, v) + \langle g, v \rangle_{\partial\Omega} \quad \forall v \in H^1(\Omega) \tag{2.23}$$

The structure of the two above problems is similar with the equation (2.18), a positive definite term and a lower order shift term. The existence of a unique solution for such problems follows from the Fredholm alternative and a uniqueness proof, see e.g. [24, 21]. To demonstrate the similarities between the problem (2.18) and the Helmholtz equation, we will shortly present the existence and uniqueness proof for the Helmholz equation with homogenous Dirichlet boundary conditions. It is easy to see, that the exactly same techniques can be used for analysis of the equation (2.18). The only difference is that different norms and spaces are involved, see [24].

To apply the Fredholm alternative, the weak problem (2.22) is cast into an operator equation,

$$(\mathcal{K} + \mathcal{I}) u = \mathcal{F} \tag{2.24}$$

where $\mathcal{K}$ is an compact operator $\mathcal{K} : L^2(\Omega) \to L^2(\Omega)$ and $\mathcal{F} \in L^2(\Omega)$. The Fredholm alternative states that

Theorem 2.1.1 (Fredholm alternative)**.** *Let $\mathcal{K} : H \to H$ be a compact linear operator and H be a Hilbert space. Then either*
(i) the equation $(\mathcal{I} + \mathcal{K})\, u = \mathcal{F}$ has a unique solution for each $\mathcal{F} \in H$
or
(ii) the equation $(\mathcal{I} + \mathcal{K})\, u = 0$ has solutions $u \neq 0$.

A proof can be found in [13]. In the following denote,

$$a(u, v) = (\nabla u, \nabla v) - \kappa^2 (u, v) + \mathrm{i}\sigma(u, v).$$

The bilinear form $a(\cdot, \cdot)$ satisfies

$$\Re a(u, u) \geq |u|_1^2 - \kappa^2 \|u\|_0^2 \quad \forall u \in H_0^1(\Omega). \tag{2.25}$$

To define the operator $\mathcal{K}$, we first need to define a new bilinear form

$$a_+(u,v) = a(u,v) + (1+\kappa^2)(u,v). \tag{2.26}$$

The new bilinear form satisfies

$$\Re a_+(u,u) \geq \|u\|_1^2,$$

i.e. it is coercive. With the help of this bilinear form, the original problem (2.22) can be written as

$$a_+(u,v) - (1+\kappa^2)(u,v) = (f,v)$$

From the operator equation (2.24), it follows that $u = \mathcal{F} - \mathcal{K}u$. Using this equation for the first term above leads to

$$a_+(\mathcal{F} - \mathcal{K}u, v) - (1+\kappa^2)(u,v) = (f,v)$$

So, we can define $\mathcal{K}$ as: For $u \in L^2(\Omega)$ find $\mathcal{K}u \in H^1(\Omega)$such that

$$a_+(\mathcal{K}u, v) = -(1+\kappa^2)(u,v) \quad \forall v \in H^1(\Omega)$$

and $\mathcal{F}$ as: Find $\mathcal{F}u \in H^1(\Omega)$ such that

$$a_+(\mathcal{F}, v) = (f,v) \qquad \forall v \in H^1(\Omega).$$

As the bilinear form $a_+(\cdot,\cdot)$ is coercive and bounded, the problems defining $\mathcal{K}$ and $\mathcal{F}$ have a unique solution by the Lax-Milgram Lemma, see [13, 21]. In addition, from the definition of operator $\mathcal{K}$, it immediately follows that

$$\|\mathcal{K}u\|_1 \leq C\|u\|_0.$$

The compactness of the operator $\mathcal{K}$ follows from the boundedness and the compact embedding of $H^1(\Omega)$ to the space $L^2(\Omega)$. The same embedding holds between X_0 and $L^2(\Omega)$, see [24]

The uniqueness of the solution is established easily: Let u_1 and $u_2$be solutions to (2.22). Then there holds

$$\Im a(u_1 - u_2, u_1 - u_2) = 0$$

This is

$$\sigma \|u_1 - u_2\|_0^2 = 0,$$

so that $u_1 = u_2$. When the parameter $\sigma$is a function or different boundary conditions are be posed, the uniqueness result follows from the unique continuation principle, see [20]. Identical results are used also to show the uniqueness of the solution to Maxwell's equations, see [24].

The existence of a unique solution follows now from the Fredholm alternative. The Fredholm alternative does not state anything about the dependence of the solution u on the load function f and the parameters σ, κ. Establishing such a stability estimate is important for the convergence analysis of the finite element method for the model problem. The following Theorem is a simplification of Theorem 2 in Publication I.

Theorem 2.1.2. *Let Ω be a convex domain, $f \in L^2(\Omega)$, $\kappa \in \mathbb{R}$, $\sigma \in \mathbb{R}$ and let u be the weak solution to (2.22). Then there exist a constant $C > 0$, independent on κ and σ, such that*

$$|u|_2 \leq C \left(1 + \frac{\kappa^2}{\sigma}\right) \|f\|_0.$$

This Theorem is proved by solving the problem (2.22) by using eigenbasis of the Laplace operator. Then a Poisson problem is formed for the solution and elliptic regularity theory is applied to obtain the desired estimate. Similar results for the problem with absorbing boundary conditions, equation (2.23), are given e.g. in [22, 23]. In their simplest form, when $g = 0$, they state that

$$|u|_2 \leq C\kappa \|f\|_0.$$

where the constant $C > 0$ is independent on κ. The results in [23] are obtained in different fashion from Theorem 2.1.2 and are more detailed. They also provide tools for analyzing convergence of higher order schemes for the Helmholtz equation. The proof given in [23] is based on splitting the solution to two parts, an analytic function and a function with limited regularity. Then κ-explicit bounds are obtained separately for the two terms. This technique is more general compared to the one used in Publication I and it can be applied to variety of different boundary conditions.

As the stability result given in Theorem 2.1.2 is important for the κ-explicit convergence analysis of numerical methods for Helmholtz equation, we will illustrate it with a numerical example. We consider the unit

square $\Omega = (0,1)^2$ and problem with homogeneous Dirichlet boundary conditions. In this case, a series solution is obtained as

$$u(x,y) = \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} \alpha_{nm}\varphi_{nm}(x,y),$$

where

$$\varphi_{nm}(x,y) = \sin \pi n x \sin \pi m y$$

and

$$\alpha_{nm} = \frac{4(f,\varphi_{nm})}{(\pi n)^2 + (\pi m)^2 - \kappa^2 + \mathrm{i}\sigma}.$$

In this setting, the semi-norm $|u|_2$ and norm $\|f\|_0$ are both easy to compute. This allows us to study the stability constant by computing the ratio $\frac{|u|_2}{\|f\|_0}$ for different right hand sides.

For $f = 1$, we have

$$a_{nm} = \begin{cases} \frac{8}{\pi^2 nm}\frac{1}{(\pi n)^2+(\pi m)^2-\kappa^2+\mathrm{i}\sigma} & \text{when } n \text{ and } m \text{ even} \\ 0 & \text{otherwise.} \end{cases}$$

By analyzing the series coefficients it is easy to see that the stability constant behaves as $O(\kappa\sigma^{-1})$. A numerical simulation was performed by approximating the series with the terms such that $n, m \leq 100$. Such an approximation to the ratio $\frac{|u|_2}{\|f\|_0}$ is presented in Figure 2.1. The resonant frequencies of the problem with $\sigma = 0$ are responsible for the spikes visible in the graph.

The predicted worst case behavior can be obtained for example, by setting

$$a_{nm} = \begin{cases} 1 & \text{when } n \leq 100 \text{ and } m \leq 100 \\ 0 & \text{otherwise.} \end{cases} \tag{2.27}$$

The ratio $\frac{|u|_2}{\|f\|_0}$ is visualized in Figure 2.2 as a function ofκ. One can clearly observe the predicted second order growth.

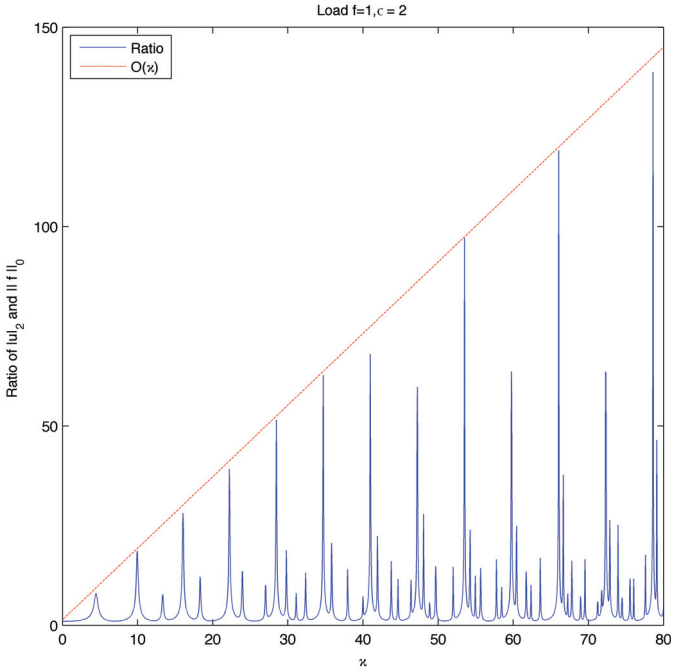

Figure 2.1. The ratio $\frac{|u|_2}{\|f\|_0}$ as a function of κ when $f = 1$ and $\sigma = 2$. The $O(\kappa)$ - behavior is clearly visible.

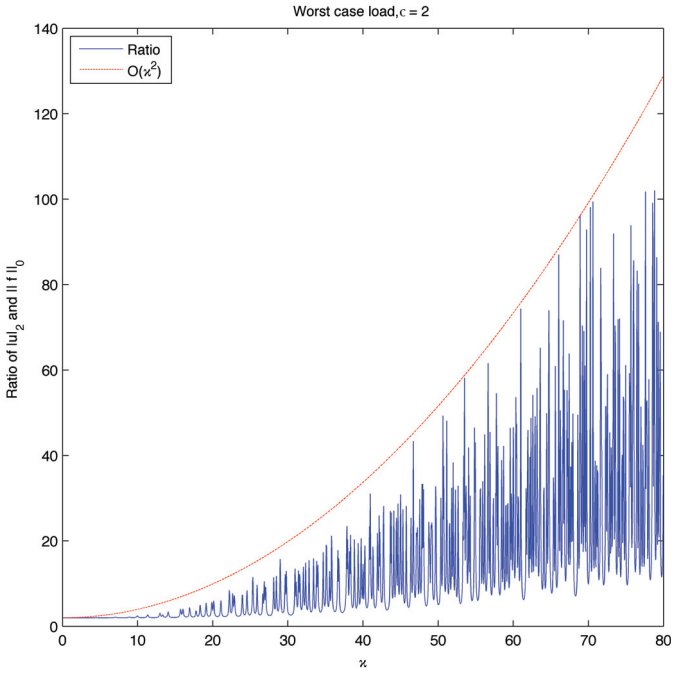

Figure 2.2. The ratio $\frac{|u|_2}{\|f\|_0}$ as a function of κ when $\sigma = 2$ and f is chosen such that the series coefficients are as in equation (2.27). The $O(\kappa^2)$ - behavior is clearly visible.

.

3. Numerical simulation of wave type phenomenon

Several different numerical methods are suitable for discretization of wave type partial differential equations. The discretization can be done using, for example, finite difference methods, boundary element method (BEM) or the finite element method. Each of these methods is best suited for certain simulations.

Due to availability of simple time-stepping schemes, the finite difference methods are very popular in time-domain simulations. These methods suffer from problems in modeling of devices with complex geometries, see e.g. [30]. The boundary element method, see [27, 21], requires a grid at all material interfaces, so it is best suited when modeling large areas of homogenous media. As unbounded domains can be easily simulated with BEM, it is widely applied for numerical solution of scattering problems.

In this thesis, we will consider numerical simulations done with the finite element method (FEM). The main benefits of FEM compared to other discretization methods are the easy handling of complex geometries and non-homogenous material parameters. The non-homogenous material parameters are encountered, for example, when non-linear medium is modeled. Typically, the linear problems solved as part of the iterative solution process of the non-linear equations have material parameters depending on the previous iterate. An example of such a situation is the numerical simulation of electrical machines constructed from steel, which is a highly non-linear material. In addition, the finite element method has a solid mathematical background. Especially this makes the mathematical treatment of different phenomenon related to numerical simulation of wave type problems using FEM possible. Good introductory texts to FEM are e.g., [2, 19].

In the finite element method, a weak form of the partial differential equation is solved approximately in a finite dimensional space V_h. The

approximate problem is: Find $u_h \in V_h$ such that

$$a(u_h, v_h) = (f, v_h) \quad \forall v \in V_h. \tag{3.1}$$

where $(\cdot, \cdot)$ is the $L^2(\Omega)$ inner product and $a(\cdot, \cdot)$ is a bilinear form related to the weak form of the PDE.

The finite element method is a systematic way to construct the subspace V_h and to assembly the matrix equation related to problem (3.1). The finite element space V_h is connected to a partition of the domain into smaller elements, for example triangles, quadrilaterals, or tetrahedrons.

We will consider the space of piecewise linear basis functions,

$$V_h := \left\{ v \in H_0^1(\Omega) \;\middle|\; v \in P_1(K) \quad \forall \quad K \in \mathcal{T}_h \right\}. \tag{3.2}$$

where $\mathcal{T}_h$ is the partition of the domain into triangular or tetrahedral elements. The parameter h is defined as the diameter of the smallest sphere containing any element of $\mathcal{T}_h$. This space is suitable for the discretization of the Helmholtz equation. A different discretization space is required for the vector-wave equation (2.14), see e.g. Publication IV.

The finite element space V_h is spanned by a set of basis functions $V_h = \operatorname{span}\{\varphi_i\}$. Each finite element function $v \in V_h$ is related to a vector of coefficients $\boldsymbol{x}_v \in \mathbb{R}^n$ via

$$v = \sum_{i=1}^{n} (\boldsymbol{x_v})_i \varphi_i. \tag{3.3}$$

The matrix equation arising form the finite dimensional problem (3.1) is

$$A\boldsymbol{x} = \boldsymbol{b}, \tag{3.4}$$

where $A_{i,j} = a(\varphi_h, \varphi_i)$ and $\boldsymbol{b}_i = (f, \varphi_i)$. In finite element simulations, the two main tasks are to construct the matrix A and to solve the linear system (3.4).

3.1 Convergence analysis and the high-frequency problem

One of the main difficulties in the numerical simulation of wave-type equations is the high number of basis functions required to resolve the solution, see [32]. To demonstrate this phenomenon, we consider the Helmholtz equation (2.19) with absorbing boundary conditions (2.21). The

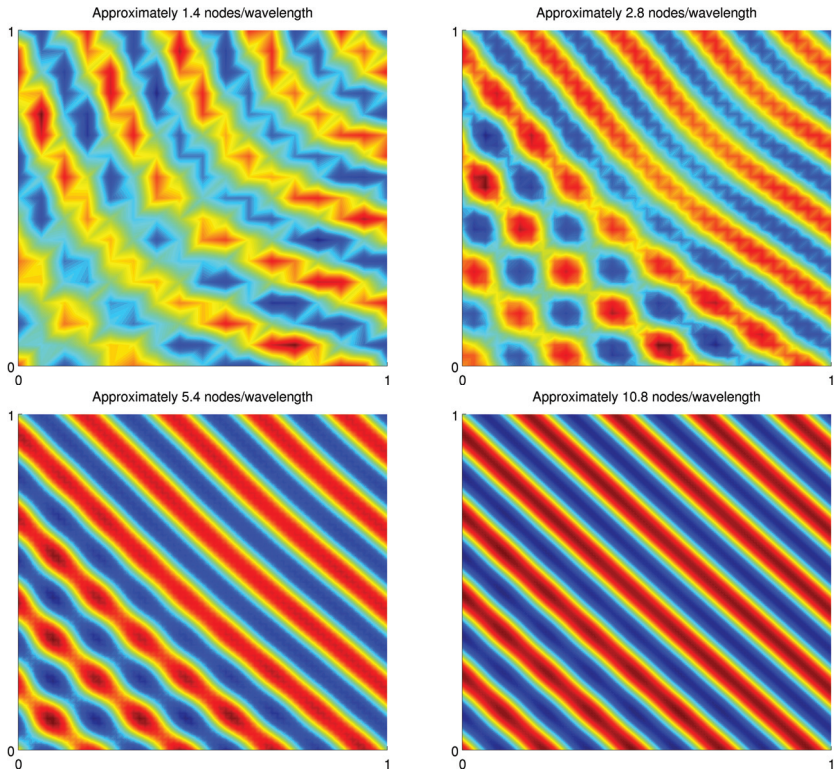

Figure 3.1. The figures show the finite element solution to problem (2.23) with $\kappa = 12\pi$ computed on a series of refining triangulations. The mesh is coarsest in the upper left and finest in the lower right corner. A sufficiently fine mesh is required, before the finite element approximation resembles the exact solution

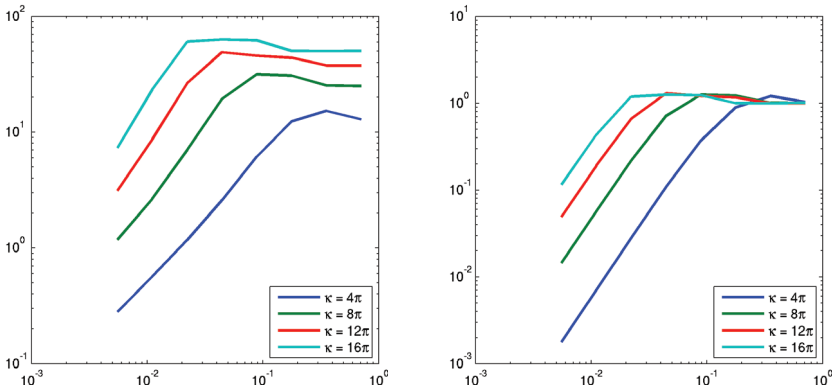

Figure 3.2. The left figure shows the error in $H^1(\Omega)$-seminorm as a function of the mesh size h and the right figure the error in $L^2(\Omega)$-norm as a function of the mesh size h. The convergence begins after a certain threshold mesh size is reached.

domain Ω is chosen as the unit square $\Omega = (0,1)^2$ and the function g is such that the exact solution is

$$u(\boldsymbol{x}) = e^{-\mathrm{i}\kappa\boldsymbol{\xi}\cdot\boldsymbol{x}}$$

where $\boldsymbol{\xi} = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\end{bmatrix}$.

The problem is solved using a triangular mesh and linear basis functions. The high-frequency problem can be easily understood based on the results show in Figures 3.1 and 3.2. Before the finite element solution visually resembles the exact solution, a sufficiently fine mesh size is required. Based on errors in the $H^1(\Omega)$-seminorm and the $L^2(\Omega)$-norm shown in Figure 3.2, the finite element approximation does not have a connection to the exact solution before a threshold mesh size is reached. Based on the results, the threshold mesh size tends to zero when the parameter κ grows.

The high-frequency problem refers exactly to the observed phenomenon: a threshold mesh size, tending to zero with growing parameter κ, is required before the finite element solution resembles the exact solution. The natural question is what is the connection between the required mesh size and κ. For our model problem, this depends on material parameters and boundary conditions.

A mathematical analysis of the connection between the threshold mesh size and κ is given in Babuška and Ihlenburg, [17, 18], for the Helmholtz equation (2.19) with absorbing boundary conditions (2.21) posed in a one dimensional domain. The analysis is divided into two parts, pre-asymtotic and asymptotic range. In the asymptotic range, the mesh has to satisfy

the constraint $\kappa^2 h \ll 1$, and the finite element error is proportional to the approximation error (i.e., quasi-optimal). Before the mesh size requirement is satisfied, i.e., in the pre-asymptotic range, an error estimate can be given if κh is sufficiently small.

In the Publication I of this thesis, the tools from the asymptotic error analysis for the Helmholtz equation with homogenous Dirichlet boundary conditions are studied in connection with preconditioned iterative methods. We will give here a simplified error estimate for the problem with homogenous Dirichlet boundary conditions using the tools of Publication I. The applied techniques are same as in [29, 17, 18].

The finite element error estimates are classically derived by relating the error in the finite element approximation to error in the interpolant of the exact solution. A convergence estimate then follows from the properties of the interpolant. For coercive problems, the interpolation and the finite element approximation errors are related to each other via Cea's Lemma, stating that

$$\|u - u_h\|_1 \leq C \inf_{v_h \in V} \|u - v_h\|_1,$$

where $C > 0$ is a positive constant, u the exact solution and u_h the finite element approximation from the space V_h. The Cea's Lemma follows from the Galerkin orthogonality property

$$a(u - u_h, v_h) = 0 \qquad \forall v_h \in V_h \tag{3.5}$$

and coercivity of the bilinear form.

$$\Re a(v, v) \geq \alpha \|u\|_1^2 \qquad \forall v \in V, \tag{3.6}$$

where $\alpha > 0$ is a positive constant. The Galerkin orthogonality holds also for the Helmholtz equation, but the coercivity property does not. Hence, the Cea's lemma cannot be directly applied. However, the coercivity will hold in a weaker sense,

$$\Re a(u - u_h, u - u_h) \geq \alpha \|u - u_h\|_1^2 \tag{3.7}$$

when the mesh size is sufficiently small. It turns out, that the above property is enough for relating the finite element approximation error to the interpolation error for the Helmholtz equation.

The bilinear form of our model problem satisfies

$$\Re a(u-u_h, u-u_h) = |u-u_h|_1^2 - \kappa^2 \|u-u_h\|_0^2 .$$

Hence, obtaining the property (3.7) requires us to relate the discretization error in the $H^1(\Omega)$-norm to error in the $L^2(\Omega)$-norm. A suitable result follows from the duality argument. In order for all constants to be independent on κ, it is important to know the κ-dependency of the stability estimate. In our case, the stability estimate is

$$|u|_2 \leq C(1+\frac{\kappa^2}{\sigma})\|f\|_0.$$

In the following analysis, we will denote $C_S = C(1+\frac{\kappa^2}{\sigma})$. The two following Theorems are simplifications of the Lemmas 4 and 5 from Publication I.

Theorem 3.1.1. *Let $u \in V$ be the weak solution to problem (2.22) and $u \in V_h$ its finite element approximation. In addition, let the mesh size h be such that*

$$1 - CC_S(\kappa^2+\sigma)h^2 > 0$$

Then there exists a constant $C > 0$, independent of κ,σ, h, and C_S, such that

$$\|u-u_h\|_0 \leq \frac{CC_S h}{1-CC_S(\kappa^2+\sigma)h^2}|u-u_h|_1$$

Theorem 3.1.2. *Let $u \in V$ be the weak solution to problem (2.22) and $u \in V_h$ its finite element approximation. In addition, let the mesh size $h$be such that*

$$C_S(\kappa^2+\sigma)h^2 \ll 1 \quad \textit{and} \quad \kappa^2 C_S^2 H^2 \ll 1$$

Then there exists a positive constant $\alpha > 0$, independent of κ,σ, h, and C_S, such that

$$\Re a(u-u_h, u-u_h) > \alpha \|u-u_h\|_1^2. \tag{3.8}$$

The requirement for the mesh size to be sufficiently small arises from the two above theorems. For our problem, the stability constant is

$$C_S = 1 + \frac{\kappa^2}{\sigma},$$

hence, if σ stays constant the mesh size should be such that the term

$$\frac{\kappa^3}{\sigma}h \ll 1. \tag{3.9}$$

Thus, in the worst case the mesh size is related to the third power of κ. When the mesh size satisfies the requirement (3.9) an error estimate follows from the Theorems 3.1.1 and 3.1.2. Using the property (3.7) and the Galerkin orthogonality (3.5) gives

$$\alpha\|u-u_h\|_1^2 \leq \Re a(u-u_h,u-u_h) = \Re a(u-\pi u_h,u-u_h).$$

in which $\pi u_h \in V_h$ is the nodal interpolant of u. The last term above can be estimated as

$$\Re a(u-\pi_h u,u-u_h) \leq |u-\pi_h u|_1\,|u-u_h|_1 + (\sigma+\kappa^2)\,\|u-\pi_h u\|_0\,\|u-u_h\|_0$$

Applying Theorem 3.1.1 to the last term yields

$$\Re a(u-\pi_h u,u-u_h) \leq |u-\pi_h u|_1\,|u-u_h|_1 + CC_S h(\sigma+\kappa^2)\,\|u-\pi_h\|_0\,|u-u_h|_1\,.$$

Dividing with $\|u-u_h\|_1$ gives the convergence estimate

$$\|u-u_h\|_1 \leq |u-\pi_h u|_1 + CC_S h(\sigma+\kappa^2)\,\|u-\pi_h u\|_0$$

which is valid only, if the assumption made on the mesh size h in Theorems 3.1.1 and 3.1.2 are satisfied. Using standard interpolation error estimates, see [2],

$$\|u-\pi_h u\|_0 \leq ch^2|u|_2 \quad \text{and} \quad |u-\pi_h u|_1 \leq ch|u|_2$$

and the stability estimate gives

$$\|u-u_h\|_1 \leq (C+CC_S k^2h^2(\sigma+\kappa^2))C_S h\|f\|_0.$$

This error estimate was obtained under the assumption that the terms $\frac{\kappa^2}{h}$ and $\frac{\kappa^3}{\sigma}h$ are small. This is a very strong requirement for the mesh size. The practical implication is that solving the problem for high-frequencies leads to extremely large linear systems.

The convergence of finite element methods for time-harmonic Maxwell's equations is similar to convergence for the Helmholtz equation. For ex-

ample, [25] presents a simple derivation of an error estimate for the finite element approximation of vector wave equation (2.14) in lossless media in a simply connected Lipschitz polyhedron with PEC boundary conditions, under the assumption that the wave number is not a resonant frequency of the problem. This bound states that there exist a constant $C(\kappa) > 0$ such that

$$\|\boldsymbol{E} - \boldsymbol{E}_h\|_{curl} \leq \frac{1}{1 - C(\kappa)h^{1/2+\delta}} \inf_{\boldsymbol{v_h} \in X_h} \|\boldsymbol{E} - \boldsymbol{v_h}\|_{curl} \tag{3.10}$$

for sufficiently small h. Here is X_h is the space of lowest order Nédélec elements, $\boldsymbol{E}_h \in X_h$ the finite element approximation, and $\delta > 0$ a parameter depending on the regularity of the solution. This bound states that as in the case of Helmholtz equation, a threshold mesh size is required before the finite element method converges. The threshold mesh size depends on the wave number κ via estimate similar to that presented in Theorem 2.1.2. Such estimates are not used in [25], leading to implicit κ-dependency. The techniques applied in [25] are similar to ones used for Helmholtz equation, but the nullspace of the curl-operator adds an additional layer of complexity to the analysis.

Several different strategies have been explored to overcome the high-frequency problem. The use of high-order basis functions is one possible strategy to battle the high-frequency problem [16, 24]. Error estimates given in [18] imply that increasing the order of the discretization is more economical than using a finer mesh. Very high-order discretizations are studied in Publication III. An alternative strategy is to look for basis functions that are more suitable to approximate wave type solutions. For example, plane wave basis functions are used in the ultra weak variational formulation, see [4, 5].

As the size of the linear system increases when the frequency grows, one approach is simply to accept the difficulties in the discretization and try to develop more efficient solvers for the linear system. Currently, computational work required to solve the linear system increases quickly when the frequency grows. More efficient preconditioners are studied in an effort to eliminate this behavior, see e.g., [7, 9, 31, 11, 10, 12] .

As a priori error estimates do not give reliable information on the convergence of the FEM-approximation, the role of a posterior error estimation becomes important. A posteriori error estimates also have a crucial role in deciding how the approximate solution could be improved while keeping the number of degrees of freedom as small as possible. Such esti-

mates are studied for Maxwell's equations in Publication IV.

3.2 Solution of the linear system

The finite element discretization of the Helmholtz equation leads to the linear system

$$A\boldsymbol{x} = \boldsymbol{b}. \tag{3.11}$$

where $A \in \mathbb{C}^{n\times n}$ and $\boldsymbol{x}, \boldsymbol{b} \in \mathbb{C}^n$.

The solution of the linear system is often the most demanding part in the numerical simulation step. When solving the linear system, two factors have to be taken into account, the computational time and the amount of memory required in the solution process. Both of these factors depend on the properties of the matrix A.

The matrix A for the finite element discretizations of the Helmholtz equation is indefinite, which makes the system (3.11) difficult to solve. This is due to lack of efficient preconditioners and large computational cost of suitable iterative methods. In addition, if material parameters or absorbing boundary conditions are present in the model, the matrix $A$is also non-normal. This means that it is not unitary diagonalizable, which has to be taken into account in evaluating different solution methods.

Solution techniques for linear systems are divided into two groups, direct and iterative methods. The direct methods are usually based on transforming the matrix A to an easily to solvable form. For example, the Cholesky decomposition for positive definite systems decomposes the matrix as product of lower triangular matrix and its transpose. Similarly, the Gaussian elimination transforms the system into a lower triangular form. Problems related to triangular matrices are solved very efficiently using back substitution.

Direct methods require the matrix to be constructed into the memory of the computer. The linear systems arising from finite element discretizations are very large and sparse. When sparse linear systems are solved with direct methods, some zeros in the sparse matrix will be transformed to non-zeros. This effect is called fill-in and it is reduced by ordering schemes, which try to eliminate the number of new non-zeros during the solution process. Due to fill-in, using direct methods for wave problems is limited by the available memory. This is the case especially in three

dimensional domains.

The second strategy for solution of linear systems are iterative methods. The iterative methods construct a sequence of approximate solutions converging towards the exact solution. These methods do not require the matrix to be constructed explicitly into the memory, instead it is sufficient just to implement the operation $A\boldsymbol{x}$. Hence, even if the linear system is too large to fit into the memory of the computer, it is still possible to solve the problem using iterative methods. The iterative methods for wave type problems are typically constrained by the long computational times, not the amount of memory required in the solution process.

The convergence of the iterative methods is dependent on the properties of the matrix A. As we will later see, the finite element discretizations of the Helmholtz equation lead to systems with the required number of iterations rapidly increasing when the frequency grows or the mesh size tends to zero. As each iteration is computationally quite expensive, such a behavior is not desirable. The remedy comes from translating the system (3.11) into a new one, for example as

$$AP\tilde{\boldsymbol{x}} = \boldsymbol{b}, \qquad \boldsymbol{x} = P\tilde{\boldsymbol{x}}.$$

This is called preconditioning. The basic idea of preconditioning is to transfer the linear system into a new one, with better iterative properties compared to (3.11). Naturally, the transformation should be inexpensive to compute. Finding a good preconditioner is often the most difficult step in iteratively solving the problem (3.11).

Preconditioners can be constructed from an algebraic viewpoint or by taking advantage of the properties of the underlying differential equation. The algebraic viewpoint usually leads to preconditioners applicable to a wide range of different problems as black-box methods. Examples of algebraic preconditioners are incomplete decompositions, such as the incomplete Cholesky or incomplete LU decomposition. The preconditioners based on the properties of the differential equation are typically more efficient but specific to a certain problems.

The current trend for solving the linear system related to time-harmonic wave problems is to use a preconditioner together with a suitable iterative method. The indefiniteness of the matrix A limits the selection of possible iterative methods. The most common choices are Krylov subspace methods, e.g. GMRES, BiCGStab, etc. (see [14, 28]). From these methods, the convergence properties are well understood only for the GMRES method.

The GMRES method is computationally a very expensive solution technique. This is due to the orthogonalization process used in the algorithm. The GMRES method stores vectors from each step into the memory and performs operations using the stored vectors on each iteration. This translates to reasonable number of GMRES iterations to be counted in tens, not in hundreds.

3.2.1 Preconditioned GMRES method

The main aim in studying preconditioned iterative methods is to understand how the preconditioner changes the iterative properties of the linear system. In practice, this question is answered by using a convergence estimate relating the properties of the matrix A to the number of iterations required to solve the problem. The convergence of GMRES is related to the matrix A as

$$|\boldsymbol{r}_i| = \min_{p \in \tilde{P}_i} |p(A)\boldsymbol{r}_0|, \tag{3.12}$$

in which $\boldsymbol{r}_i$ is the residual at step i and $\tilde{P}_i$ is the set of monic polynomials of order i. The minimization problem (3.12) is very difficult to solve and thus it is not a practical measure of the GMRES convergence rate. More useful bounds have been derived from (3.12) in several alternative ways, depending on the properties of matrix A, see [28, 14, 8].

Selecting the convergence criterion depends on the spectral properties of matrix A. The matrix arising from the model problem is non-normal, i.e.,

$$AA^* \neq A^*A.$$

This means, that the matrix is A is not unitary diagonalizable. For non-normal matrices, the convergence of GMRES does not depend solely on the eigenvalues, but also on the eigenvectors, see [14, 8]. In the Publication I, a field of values (FOV) based convergence criterion for GMRES is used to study the iterative properties of the GMRES method. The field of values of matrix A is defined as the set

$$\mathcal{F}(A) = \left\{ \frac{\boldsymbol{x}^* A \boldsymbol{x}}{\boldsymbol{x}^* \boldsymbol{x}} \;\middle|\; \boldsymbol{x} \in \mathbb{C}^n,\ \boldsymbol{x} \neq 0 \right\}. \tag{3.13}$$

The location of this set in the complex plane is related to the convergence properties of GMRES for matrix A. A simple estimate is given in [14], let

$D = \{ z \in \mathbb{C} \mid |z - c| \leq s \}$ be a disc containing the FOV, but not the origin. Then, we have the convergence estimate

$$|\boldsymbol{r}_i| \leq \left(\frac{s}{|c|} \right)^i |\boldsymbol{r}_0|. \tag{3.14}$$

To demonstrate the difficulties in solving the Helmholtz equation using iterative methods, we give a quick FOV based convergence analysis for the non-preconditioned system. We will again consider the Helmholtz equation with homogenous Dirichlet boundary conditions and the first order finite element space (3.2). The presented analysis uses techniques from Publication I. The aim in the analysis is to give bounds for the location of the FOV set in the complex plane, which leads to convergence estimate via (3.14).

The field of values is related to the bilinear form $a(\cdot, \cdot)$ as

$$\frac{\boldsymbol{x}_u^* A \boldsymbol{x}_u}{\boldsymbol{x}_u^* \boldsymbol{x}_u} = \frac{a(u,u)}{\boldsymbol{x}_u^* \boldsymbol{x}_u}.$$

Here we have used the notation $\boldsymbol{x}_u$ for the vector of coefficients of the finite element function u, i.e.

$$u(\boldsymbol{x}) = \sum_{i=1}^{N} (\boldsymbol{x}_u)_i \varphi_i(\boldsymbol{x}),$$

where $\varphi_i, i = 1, \ldots, N$ are the basisfunctions of the finite element space V_h. This notation will also be used in the following analysis.

Taking an imaginary part of bilinear form gives

$$\Im a(u,u) = \sigma \|u\|_0^2 \tag{3.15}$$

and real part

$$\Re a(u,u) = |\nabla u|_1^2 - \kappa^2 \|u\|_0^2. \tag{3.16}$$

The Euclidian norm of the coefficient vector and the $L^2(\Omega)$-norm of the corresponding function are related as

$$c h^d \boldsymbol{x}_u^* \boldsymbol{x}_u \leq \|u\|_0^2 \leq C h^d \boldsymbol{x}_u^* \boldsymbol{x}_u, \tag{3.17}$$

where $c > 0$ and $C > 0$ are positive constants and d is the spatial dimension, see e.g. [2]. Using this identity and equation (3.15) gives

$$c\sigma h^d \leq \Im \frac{a(u,u)}{\boldsymbol{x}_u^* \boldsymbol{x}_u} \leq C\sigma h^d.$$

The inverse inequality, see e.g. [2], states that

$$|v_h|_1 \leq Ch^{-1}\|v_h\|_0 \quad \forall v_h \in V_h$$

where $C > 0$ is a positive constant independent of h. Estimating (3.16) with Poincaré-Friedrichs and the inverse inequality gives

$$(C - \kappa^2)\|u\|_0^2 \leq \Re a(u,u) \leq Ch^{-2}\|u\|_0^2 \tag{3.18}$$

combining this with equation (3.17) yields

$$ch^d(C - \kappa^2) \leq \frac{\Re a(u,u)}{\boldsymbol{x}_u^* \boldsymbol{x}_u} \leq Ch^{d-2}. \tag{3.19}$$

These estimates state that the FOV is located inside a rectangle,

$$\mathcal{F}(A) \subset \left[ch^d(C - \kappa^2), Ch^{d-2}\right] \times \left[ch^d, Ch^d\right]. \tag{3.20}$$

A refined estimate can be obtained in the spirit of Publication I. Let $z \in \mathcal{F}(A)$. Then there exists a function u such that

$$\Im z = \frac{a(u,u)}{\boldsymbol{x}_u^* \boldsymbol{x}_u}.$$

This leads to the equality

$$\|u\|_0^2 = \frac{\Im z \boldsymbol{x}_u^* \boldsymbol{x}_u}{\sigma}. \tag{3.21}$$

The real part of z satisfies

$$\Re z = \Re \frac{a(u,u)}{\boldsymbol{x}_u^* \boldsymbol{x}_u} = \frac{|\nabla u|_1^2 - \kappa^2 \|u\|_0^2}{\boldsymbol{x}_u^* \boldsymbol{x}_u}.$$

Combining this with equation (3.21) yields

$$\Re z = \frac{|\nabla u|_1^2}{\boldsymbol{x}_u^* \boldsymbol{x}_u} - \frac{\kappa^2}{\sigma} \Im z.$$

Thus, the FOV set is located at the intersection of the box (3.20) and the strip

$$S = \left\{ z \in \mathbb{C} \mid ch^d - \frac{\kappa^2}{\sigma}\Im z \leq \Re z \leq Ch^{d-2} - \frac{\kappa^2}{\sigma}\Im z \right\}.$$

The location of the FOV set determines the convergence of the GMRES method. As the FOV set is bounded in quite a complicated domain, we will just state that the convergence is dependent on h,κ^2, and $\frac{\kappa^2}{\sigma}$. The size of FOV set grows when the mesh size tends to zero. Growing the parameter

κ has also a major effect to the convergence. The mesh size dependency can be eliminated relatively easily, but eliminating the κ^2-dependency is considerable more difficult. Such methods are studied in Publication I.

The FOV sets computed using the same procedure as in Publication I for mesh sizes h_0,$\frac{1}{2}h_0$, and $\frac{1}{4}h_0$ are presented in Figure 3.3. The mesh size dependency of FOV is apparent from these results.

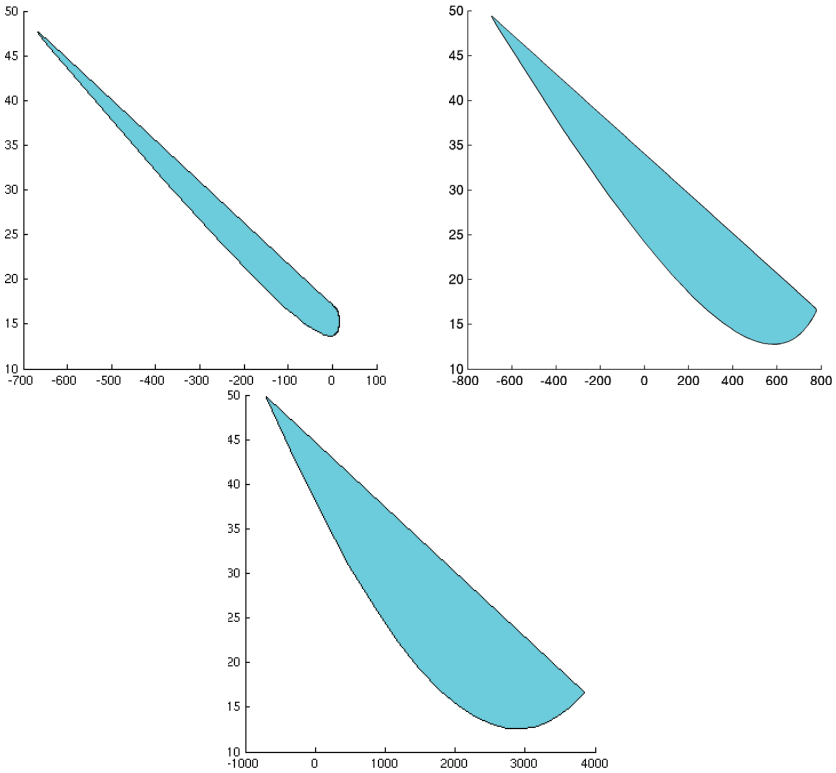

Figure 3.3. The h^2-scaled field of values sets for $\sigma = 100$, $\kappa = 12\pi$. The mesh parameter is largest in the upper left figure and smallest in the lowest figure. The mesh parameter is divided into half between the figures.

3.3 Domain decomposition methods

When the linear system (3.11) is too large to fit into the memory of a single computer or the required computational time is simply too long, domain decomposition methods can be used. The idea in domain decomposition methods is to divide the domain Ω into smaller subdomains. The subdomains can be distributed to a several computers, thus relaxing the memory requirements. To solve the problem, a set of equations connecting

the solution at the subdomain interfaces is formulated and solved.

The domain decomposition methods for the Helmholtz equation are also used without parallel computing. If an efficient way to solve the subdomain problems exists, the problem can be posed only for the interface conditions. Such an approach is used in the ultra weak variational formulation, [4, 5], in which a plane wave basis functions are used on each element. The plane wave basis functions allow the problem at the interior of the subdomains to be solved analytically, and the resulting problem reduces to one for the interface conditions.

The benefits in reducing the original problem to one posed on the subdomain interfaces are better iterative properties and the reduction in the size of the problem. Both of these benefits are important for the Helmholtz equation.

The success of domain decomposition methods for Helmholtz equation depends on how the interface conditions are chosen. Choosing the nodal values at the interfaces as unknowns similarly as in methods for the Laplace problem leads to big difficulties in the convergence. For example, in the ultra weak variational formulation, the interface conditions

$$\frac{\partial u_1}{\partial \boldsymbol{n}} - \mathrm{i}\kappa u = \frac{\partial u_2}{\partial \boldsymbol{n}} + \mathrm{i}\kappa u_2 \tag{3.22}$$

$$\frac{\partial u_1}{\partial \boldsymbol{n}} + \mathrm{i}\kappa u = \frac{\partial u_1}{\partial \boldsymbol{n}} - \mathrm{i}\kappa u_2 \tag{3.23}$$

are imposed on the interfaces. Here u_1 is the solution on the subdomain left to the interface and u_2 right to the interface. These interface conditions lead to convergent method.

In the Publication III a domain decomposition-type method for solving the Helmholtz equation is developed. This method is based on hybridized mixed Helmholtz equation. Hybridization is a general solution strategy for mixed system and can also be used as a tool for domain decomposition, see [3]. In hybridization, additional variables are introduced to enforce the continuity conditions over subdomain interfaces. This leads to subdomain problems that are coupled via interface conditions. The subdomain problems are then solved, which leads to a new set of equations only for the interface unknowns. As for all domain decomposition methods, the interface conditions for hybridization of the mixed Helmholtz equation have to be chosen with care. In [26] the normal continuity of the flux is broken and an additional unknown is used for stability, which leads to method

with good iterative properties.

4. Concluding remarks

4.1 Publication I

In Publication I we study the solution of the linear system arising from the Helmholtz equation (2.19) with homogenous Dirichlet boundary conditions (2.20). The focus is in analyzing three preconditioners for the GMRES method. As we have discussed in Section 3.2.1, the required number GMRES iteration without a preconditioner grows rapidly when the parameter κ increases or the mesh size tends to zero.

The simplest precoditioner discussed in Publication I is the Laplace preconditioner, which eliminates the mesh size dependency from the required number of GMRES iterations. However, a κ^2-dependency still remanins. When the frequency grows, this dependency leads to rapid growth in the required number of iterations. However, as it is shown in the article, the Laplace preconditioner can be evaluated using the multigrid method making it fast to compute even for large number of unknowns.

To eliminate the κ^2-dependency from the number of iterations, a two-level preconditioner is introduced. The two-level preconditioner combines solution on a coarse grid with the Laplace preconditioner. Such a method succeeds in eliminating both the κ and the mesh size dependency from the required number of iterations. Unfortunately, the same mesh size constraints as we derived for the problem in Section 3.1 have to be satisfied also by the coarse grid mesh size, leading to preconditioner that is very expensive to evaluate.

4.2 Publication II

In Publication II, the solution of the linear system arising from mixed Poisson problem

$$\begin{aligned} -\nabla u + \boldsymbol{\sigma} &= 0, \\ \operatorname{div}(\boldsymbol{\sigma}) &= f, \end{aligned} \tag{4.1}$$

is studied. The mixed Poisson problem does not exhibit any wave type phenomenon. The motivation for studying this system was born from the desire to mathematically analyze the method presented in Publication III, which is based on hybridization of mixed Helmholtz equation given in [26].

The idea in the preconditioner presented in Publication II is to use existing preconditioners for the Poisson problem also in the solution of the mixed problem. The main benefit of this approach is the possibility to use well tested and already implemented methods also with the mixed system. The properties of the preconditioned system are analyzed and numerical test verifying its efficiency are presented.

4.3 Publication III

In Publication III, a hybridization method for solving the Helmholtz equation (2.19) with absorbing boundary conditions (2.20) is presented. As discussed in Section 3.3, the hybridization is a domain decomposition method used to reduce the mixed problem to problem posed on subdomain interfaces.

In this publication, the reduction is done on a structured rectangular grid with the help of a special very high-order polynomial basis. The basis is constructed from one dimensional eigenfunctions by taking advantage of the tensorial structure of the gird. The tensorial eigenbasis allows the problems related to the interiors of the subdomains to be solved extremely efficiently, leading to cheap reduction of the system to subdomain interfaces.

To solve the interface problem, a preconditioned iterative method is applied. Numerical examples show that the resulting solution strategy is very efficient and it is well suited for solving high-frequency problems. The downside of the method is the special structure required from the

mesh. This complicates the modeling of difficult geometries, but makes it possible to cheaply eliminate the subdomain problems.

4.4 Publication IV

In Section 3.1, a convergence estimate for a finite element method was derived under assumptions on the mesh size. Similar estimates hold also for other time-harmonic wave equations, e.g., the vector wave equation (2.14). The main value of these a priori error estimates is that they guarantee that the finite element method eventually convergences to the exact solution. However, in practice, it is difficult to determine when the finite element solution is a good approximation to the exact solution. A posteriori error estimates provide an answer to this question.

In the Publication IV, we present functional type a posteriori error estimates for the Maxwell's equations. We consider in detail the eddy-current problem and shorty the vector wave equation (2.14). The presented estimates allow guaranteed upper bounds to be computed for the finite-element discretization error.

The reader should note a possible shortcoming of the method presented in this publication. The parameter $\boldsymbol{y}^* \in H(\mathrm{curl};\Omega)$ required in the estimate is computed by approximately solving the problem: find $\boldsymbol{y}^* \in H(\mathrm{curl};\Omega)$ such that

$$(\beta^{-1}\mathrm{curl}(\boldsymbol{y}^*),\mathrm{curl}(v)) + (\boldsymbol{y}^*,\boldsymbol{v}) = (\beta^{-1/2}(\boldsymbol{f}-\beta\tilde{\boldsymbol{u}}),\mathrm{curl}(u)) + (\mathrm{curl}(\tilde{\boldsymbol{u}}),\boldsymbol{v}) \quad \forall \boldsymbol{v} \in H(\mathrm{curl};\Omega)$$

in a finite dimensional space. The load function of this problem can be badly behaving, even for a function $\boldsymbol{f} \in L^2(\Omega)$. This may affect the convergence of $\boldsymbol{y}^*$, in the worst case leading to different convergence rates for the actual error and the presented estimator. Such a behavior is not studied in Publication IV and it requires a throughout investigation.

Bibliography

[1] A. Bossavit. *Computational electromagnetism*. Electromagnetism. Academic Press Inc., San Diego, CA, 1998. Variational formulations, complementarity, edge elements.

[2] D. Braess. *Finite elements*. Cambridge University Press, Cambridge, 2007.

[3] F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods*, volume 15 of *Springer Series in Computational Mathematics*. Springer-Verlag, New York, 1991.

[4] O. Cessenat and B. Despres. Application of an ultra weak variational formulation of elliptic PDEs to the two-dimensional Helmholtz problem. *SIAM J. Numer. Anal.*, 35(1):255–299 (electronic), 1998.

[5] O. Cessenat and B. Després. Using plane waves as base functions for solving time harmonic equations with the ultra weak variational formulation. *J. Comput. Acoust.*, 11(2):227–238, 2003.

[6] M. Dauge. *Elliptic boundary value problems on corner domains*, volume 1341 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1988.

[7] H. C. Elman, O. G. Ernst, and D. P. O'Leary. A Multigrid Method Enhanced by Krylov Subspace Iteration for Discrete Helmholtz Equations. *SIAM J. Sci. Comput.*, 23:1291–1315, April 2001.

[8] M. Embree. How descriptive are GMRES convergence bounds? Technical report, Oxford University Computing Laboratory, 1999.

[9] Y. A. Erlangga. Advances in iterative methods and preconditioners for the Helmholtz equation. *Arch. Comput. Methods Eng.*, 15(1):37–66, 2008.

[10] Y. A. Erlangga, C. W. Oosterlee, and C. Vuik. A novel multigrid based preconditioner for heterogeneous Helmholtz problems. *SIAM J. Sci. Comput.*, 27(4):1471–1492, 2006.

[11] Y. A. Erlangga, C. Vuik, and C. W. Oosterlee. On a class of preconditioners for solving the Helmholtz equation. *Appl. Numer. Math.*, 50(3-4):409–425, 2004.

[12] Y. A. Erlangga, C. Vuik, and C. W. Oosterlee. Comparison of multigrid and incomplete LU shifted-Laplace preconditioners for the inhomogeneous Helmholtz equation. *Appl. Numer. Math.*, 56(5):648–666, 2006.

[13] L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.

[14] A. Greenbaum. *Iterative Methods for Solving Linear Systems*. SIAM, 1997.

[15] P. Grisvard. *Singularities in boundary value problems*, volume 22 of *Recherches en Mathématiques Appliquées [Research in Applied Mathematics]*. Masson, Paris, 1992.

[16] R. Hiptmair. Finite elements in computational electromagnetics. *Acta Numerica*, 11:237–339, January 2002.

[17] F. Ihlenburg and I. Babuška. Finite element solution of the Helmholtz equation with high wave number. I. The h-version of the FEM. *Comput. Math. Appl.*, 30(9):9–37, 1995.

[18] F. Ihlenburg and I. Babuška. Finite element solution of the Helmholtz equation with high wave number. II. The h-p version of the FEM. *SIAM J. Numer. Anal.*, 34(1):315–358, 1997.

[19] C. Johnson. *Numerical solution of partial differential equations by the finite element method*. Dover Publications Inc., Mineola, NY, 2009. Reprint of the 1987 edition.

[20] R. Leis. *Initial-boundary value problems in mathematical physics*. B. G. Teubner, Stuttgart, 1986.

[21] W. McLean. *Strongly elliptic systems and boundary integral equations*. Cambridge University Press, Cambridge, 2000.

[22] J. M. Melenk. *On Generalized Finite Element Methods*. PhD thesis, University of Maryland, 1995.

[23] J. M. Melenk and S. Sauter. Wavenumber explicit convergence analysis for Galerkin discretizations of the Helmholtz equation. *SIAM J. Numer. Anal.*, 49(3):1210–1243, 2011.

[24] P. Monk. *Finite element methods for Maxwell's equations*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2003.

[25] P. Monk. A simple proof of convergence for an edge element discretization of Maxwell's equations. In *Computational Electromagnetics*, volume 28 of *Lecture Notes in Computational Science and Engineering*, pages 127–141. Springer Berlin Heidelberg, 2003.

[26] P. Monk, J. Schöberl, and A. Sinwel. Hybridizing Raviart-Thomas Elements for the Helmholtz Equation. *Electromagnetics*, 30:149–176, 2010.

[27] J-C. Nédélec. *Acoustic and electromagnetic equations*, volume 144 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2001.

[28] Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, 2003.

[29] A. H. Schatz. An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. *Math. Comp.*, 28:959–962, 1974.

[30] A. Taflove and S. C. Hagness. *Computational electrodynamics: the finite-difference time-domain method*. Artech House Inc., Boston, MA, second edition, 2000.

[31] M. B. van Gijzen, Y. A. Erlangga, and C. Vuik. Spectral analysis of the discrete Helmholtz operator preconditioned with a shifted Laplacian. *SIAM J. Sci. Comput.*, 29(5):1942–1958, 2007.

[32] O. C. Zienkiewicz. Achievements and some unsolved problems of the finite element method. *Internat. J. Numer Model.*, 47:9–28, 2000.

Errata

Publication IV

The function $\boldsymbol{y}^*$ should be from the space $H(\mathrm{curl};\Omega)$ throughout Publication IV. The proofs of Theorems 1 and 2 are valid under this assumption. The numerical examples were performed by solving equation (6) in the whole finite element space, without imposing any boundary conditions.

Aalto-DD 88/2011

ISBN 978-952-60-4297-8 (pdf)
ISBN 978-952-60-4296-1
ISSN-L 1799-4934
ISSN 1799-4942 (pdf)
ISSN 1799-4934

Aalto University
School of Science
Department of Mathematics and Systems Analysis
www.aalto.fi

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

**DOCTORAL
DISSERTATIONS**